Latent variable modeling - Lecture 1

(*Academic year 2024-2025*).

Organization and summary

- Agenda: 8 lectures of 3 hours.
- Grading: 3h written exam.
- **Aim:** In this course, we study some latent variable models using (mainly) a Bayesian approach. We particularly focus on mixture models. All along the different encountered data problems, we provide basic description of some underlying MCMC algorithms (Gibbs, Metropolis-Hasting, ABC method, EM algorithm, Dirichlet process stick breaking, ...). Introduction to Bayesian language programming is provided.

  Summary:

  - The Bayesian paradigm:
    - some standard parametric models (Bernoulli, Normal, Gamma, regression processes...).
    - Credibility intervals, Bayesian hypothesis testing, asymptotic results, decision theory,
    - Objective and subjective prior specification.
  - Bayesian two sample problems, linear regression, hierarchical models, changepoint detection,
  - Mixture models,
  - Nonparametric models (Dirichlet Process (DP), DP mixture, Gaussian processes) for density estimation and nonparametric regression.
- **software**: R, Jags, Stan, Jasp.
- **Bibliography:**

  - Robert, C.P. (2007). The Bayesian choice. From decision-theoretic foundations to computational implementation. Springer texts in statistics, Second edition.

Today's introductory lecture:

- A primer on Bayesian statistics

Contents

1 A primer in Bayesian inference

Different paradigms for statistical inference:

- Classical or frequentist,
- Bayesian,
- Fiducial,
- ...

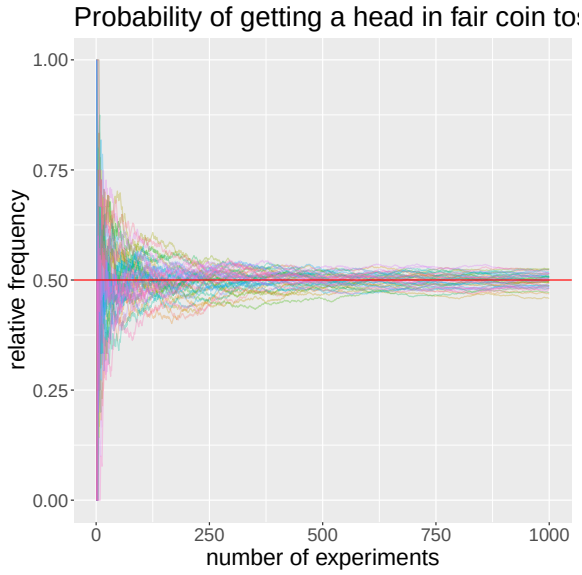
1.1 On the meanings of probability

"Probability and Proof" from Prof. Dawid's website (*www.statslab.cam.ac.uk/ apd/*). There is agreement on the math for probability, but not on the interpretation.

Different meanings:

- Statistical,
- Classical,
- Empirical or frequentist,
- Subjective,
- Logical,
- ...

Illustration of empirical probability:



As we will see, there are limitations to this definition, which serves as a basis for frequentist statistics. We will now view probability as 'logical' or 'subjective'.

R.T. Cox (1946) Probability, frequency and reasonable expectation, American Journal of Physics. vol 14, pp 1-13. Plausibility quantifier for a **rational person** in a **given context**. It should satisfy the three following properties:

1. Degrees of plausibility are represented by real numbers (the larger the number, the larger the plausibility): Notations:
   - $\bar{A}$: assertion A is false.
   - $P(A)$: plausibility of assertion A.
   - $P(AB)$: plausibility that assertions A and B are simultaneously true.
   - $P(A+B)$: plausibility that either assertion A or B are true.
   - $P(A|B)$: plausibility that A is true, knowing that B is true.
2. Reason consistently:
   - If a conclusion can be reached in more than one way, then it should lead to the same degree of plausibility.
   - All the available evidence should be used to quantify plausibility: no information can be left out.
   - Equivalent states of knowledge should give same plausibility assignments.
3. Qualitative correspondence with common sense:
   - Consider statements A and B and some contextual information H.
   - Assume that an update of H is given by H'.
   - Assume that the plausibilities are affected as follows:
     - $P(A|H') > P(A|H)$
     - $P(B|AH') = P(B|AH)$.

     Then the requested property is that:

     $$P(AB|H') > P(AB|H)$$

     More plausibility after update of the information.

Cox(1946) this gives us back the Kolmogorov axioms from probability theory.

- *Bounds:*

  $$0 \leq P(A|H) \leq 1,$$

  where A is an event, $P(A|H) = 0$ if A is impossible and $P(A|H) = 1$ if A is certain in the context H.
- *Addition rule:* If A and B are mutually exclusive (i.e. one at most can occur)

  $$P(A \cup B|H) = P(A|H) + P(B|H).$$
- *Multiplication rule:* For any events A and B,

  $$P(A \cap B|H) = P(A|B,H)P(B|H).$$

  We say that A and B are independent if $P(A \text{ and } B|H) = P(A|H)P(B|H)$ or equivalently $P(A|B,H) = P(A|H)$.

About subjectivity and context:

- All probabilities are conditional on context H
- They are **Your probabilities** for an event, not a property of the event.

- Probabilities are therefore subjective and can be given for unique events, e.g. the probability of aliens openly visiting earth in the next 10 years.
- They express **Your relationship** to the event - different stakeholders will have different information and different probabilities.

Is probability Chance or Ignorance?
We can think of two broad types (at least) of uncertainty:

- Aleatory: essentially unpredictable.
- Epistemic: due to lack of knowledge.

From subjectivist point of view, no need to worry about distinction between them, they are just uncertainties, e.g.

- What is the probability that it will rain tomorrow ?

  We often say that the probability of rain tomorrow is 3/4. Is tomorrow weather random or deterministic ? meteorologic models are complex dynamic deterministic. But uncertainties w.r.t initial conditions and chaos phenomena make these previsions uncertain and random...
- What is the height of Nigara Falls ?

  From a practical point of view, it is also natural to give probability law to non random parameters. For example, it is natural to give a probability of 0.9 that the height of Nigara Falls is in between 40 and 70m. This height is of course not random, we are just characterizing our uncertainty about its true value.

Subjectivity is of often seen as a major disadvantage, however one should not confuse subjectivity with the **difficulty to translate beliefs about an assertion A into the number $P(A|H)$**.

Inversion of probabilities is described by the **Bayes formula.** It provides a formal mechanism for learning from experience. Let us consider two events A, E, such that $P[E \neq 0]$.

$$P[A|E] = \frac{P[E|A]P[A]}{P[E|A]P[A] + P[E|A^c]P[A^c]} = \frac{P[E|A]P[A]}{P[E]},$$

where A^c denotes the complementary event.
Now let us look at the following ratio

$$\frac{P[A|E]}{P[B|E]} = \frac{P[E|A]P[A]}{P[E]} \frac{P[E]}{P[E|B]P[B]}.$$

If events A and B are equiprobable, then

$$\frac{P[A|E]}{P[B|E]} = \frac{P[E|A]}{P[E|B]}.$$

For 2 equiprobable causes, the ratio of the probabilities of the causes conditionally to the effect is the same as the ratio of the probability of the effect conditionally to the causes.

1.2 Bayesian statistical model

The specification of a Bayesian model starts with a statistical model $\mathcal{P} = \left\{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\right\}$ $d \geq 1$, fixed. The beliefs on θ are represented by a random variable $\bar{\Theta} \sim \Pi$, where Π is a **prior** probability distribution endowing the parameter space Θ.

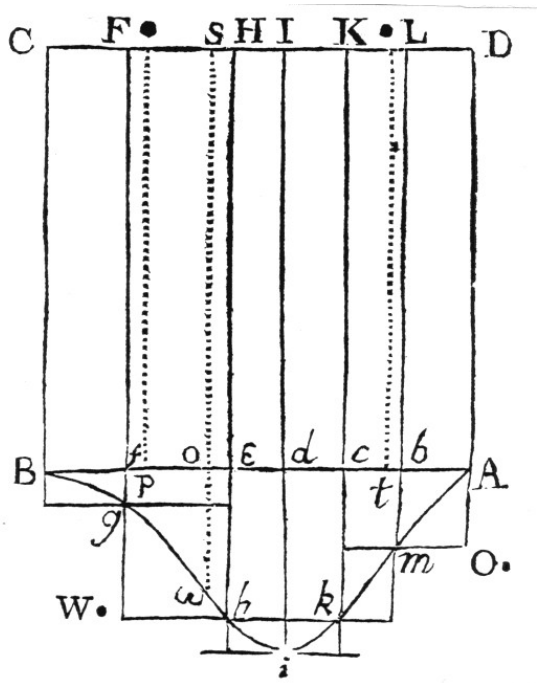
The sample $(X_1, \ldots, X_n)$ is viewed as realizations of the conditional law $X|\theta \sim P_\theta$ given that $\bar{\Theta} = \theta$.
The conditional distribution of $\bar{\Theta}|X_1, \ldots, X_n$ is the **posterior** distribution. The model is said to be dominated, when the distributions P_θ and Π admit densities $p(\cdot|\theta)$ and $\pi(\cdot)$ respectively. From now on, for simplicity of notations, all variables will be written in small letters, the context will allow understanding whether it concerns random variable or not). Using the Bayes theorem, the posterior is written as follows:

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\prod_{i=1}^n p(x_i|\theta)\pi(\theta)}{\int \prod_{i=1}^n p(x_i|\theta)\pi(\theta)d\theta}$$

Starting from the effects (observations) one try to infer on the causes (the parameters $\theta \in \Theta$) of the Data Generating Process. This is the inversion w.r.t probabilistic approach; Likelihood = inverted density! $l(\theta; x) = p(x|\theta)$.

1.2.1 Historical foundations

Bayes Billard table



Bayes, T. (1763) An Essay towards solving a problem in the Doctrine of Chances, Philosophical transactions of the Royal Society, 53, 370-418. A billiard ball is launched on a line of length 1. The location $\theta \in (0,1)$ where the ball stop is modeled by a uniform law. A second ball is launched n times from identical manner and y is the number of times that this ball is on the right of the first one. Knowing y, what can we say on θ?

$$\begin{array}{rl} \text{Likelihood} & \left\{ \begin{array}{l} y|\theta \sim Bin(n,\theta) \\ p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}. \end{array} \right. \\ \text{Prior} & \left\{ \begin{array}{l} \theta \sim U[0,1] \\ \pi(\theta) = 1_{\{\theta\in(0,1)\}}. \end{array} \right. \\ \text{Posterior} & \left\{ \begin{array}{l} \theta|y \sim Be(y+1, n-y+1) \\ \pi(\theta|y) \propto \theta^y(1-\theta)^{n-y}. \end{array} \right. \end{array}$$

$\propto$ denotes proportional up to a multiplicative constant which does not depend on θ.

```
# true parameter
theta = 0.17
# n observations
n = 60
# generate example data
set.seed(123)
x = rbinom(n = n, prob = theta, size = 1)
```

Reminder: one classical frequentist solution Use normal approximation to build a 95% confidence interval for the proportion θ:

$$\hat{\theta} \pm z_{0.975}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

```
# standard frequentist estimator
y = sum(x)
theta_hat = y / n
```

```r
# 95% confidence interval
LB = theta_hat - qnorm(0.975)* sqrt((theta_hat*(1- theta_hat))/n)
UB = theta_hat + qnorm(0.975)* sqrt((theta_hat*(1- theta_hat))/n)
```

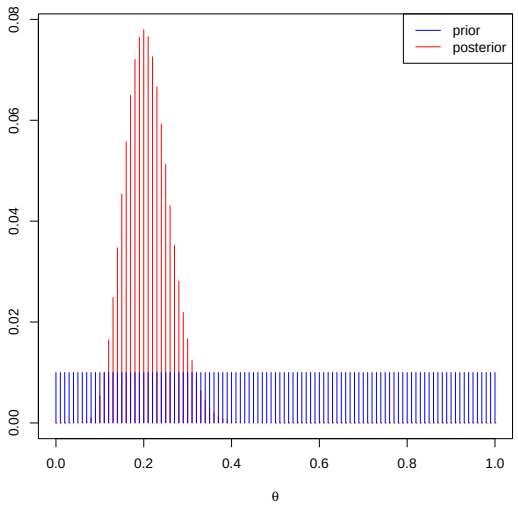
```
## theta_hat,  0.2
## 95% confidence interval,  0.0988 0.3012
```

Remark: The normal approximation deteriorates when the unknown is close to 0 or 1.

1.2.2 Bayesian approach for the Binomial model

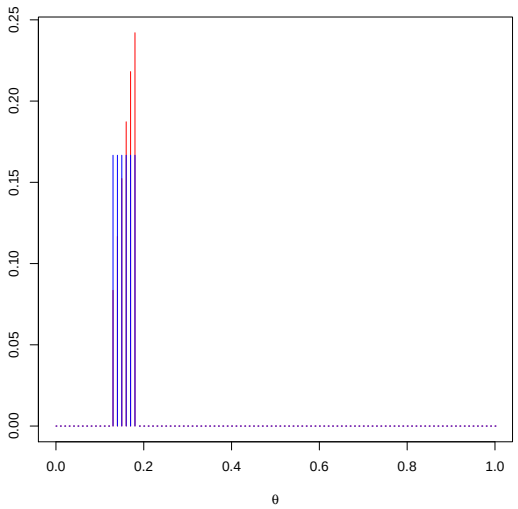
using a discrete uniform prior on θ over the set values $\{i/100; 0 \leq i \leq 100\}$.

```r
theta=seq(0,1,length = 101)
prior = rep(1/100, length = 101)
like = choose(n,y)*(theta^y)*(1-theta)^(n-y)
post = prior * like / sum(like*prior)
```



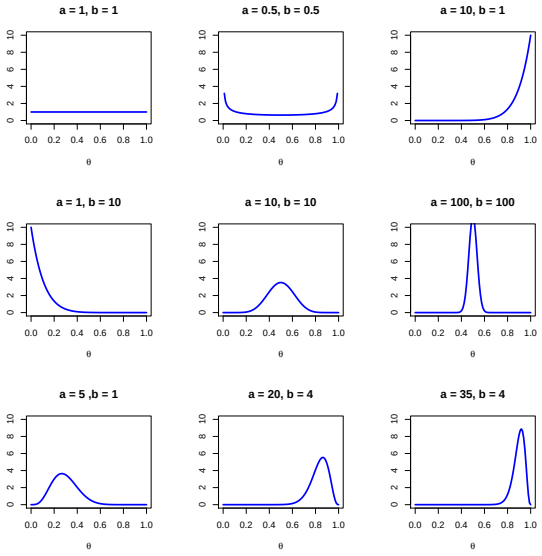

```
## prior probability of theta in [0.15;0.20]= 0.05
## posterior probability of theta in [0.15;0.20]= 0.3145219
```

using a truncated discrete uniform prior on θ over the set values $\{i/100; 14 \leq i \leq 19\}$.

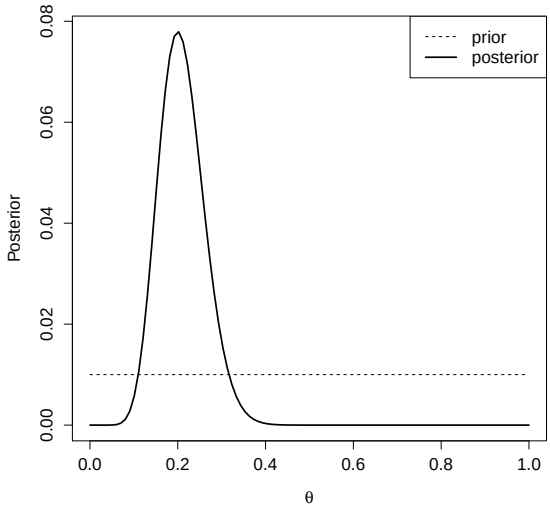


```
## prior probability of theta in [0.14;0.19] 0.8333333
## posterior probability of theta in [0.14;0.19] 0.9164723
```

using a continuous beta prior Shapes of the beta distribution




```
#set the prior distribution parameters
a  = b = 1

# posterior parameters
a_post = y + a
b_post = n - y + b

# Plot the prior, likelihood and posterior
L = 100
theta <- seq(0,1,length=L)
theta_prior <- dbeta(theta,a, b)/L
theta_post <- dbeta(theta,a_post, b_post)/L
```



1.2.3 Summarizing the information from the posterior

Definition 1.1. *Consider a parametric Bayesian statistical model for the observation x, which consists in a sampling family $\{P_\theta; \theta \in \Theta\}$ and a prior distribution Π on $\theta \in \Theta$. We define*

- *the posterior mean*

$$\bar{\theta} = \mathbb{E}_{\theta|x}(\theta) = \int_{\Theta} \theta d\Pi(\theta|x).$$

- *the posterior mode, denoted as* $\hat{\theta}_m$*: i.e., the point(s) where the posterior density is maximum,*

$$\bar{\theta}_m = \arg\max_{\theta \in \Theta} \pi(\theta|x).$$

- *the posterior variance*

$$\bar{v} = \mathbb{V}ar_{\theta|x}(\theta) = \int_{\Theta} (\theta - \bar{\theta})^2 d\Pi(\theta|x).$$

Definition 1.2. *Let* $\Theta \subset \mathbb{R}$*, and* $F_{\theta|x}$ *be the c.d.f(cumulative distribution function) of the posterior distribution* $\Pi(\theta|x)$*. Suppose* $F_{\theta|x}$ *as an inverse* $F^{-1}_{\theta|x}$*. We define the t-th posterior quantile as*

$$q_{\theta|x}(t) = F^{-1}_{\theta|x}(t).$$

Definition 1.3 (credible region)**.** *A region* C *of* Θ *is said to be a* $(1-\alpha)$ *credible region for* $\Pi(\cdot|x)$ *if and only if*

$$\Pi(\theta \in C|x) \geq 1 - \alpha.$$

Remarks:

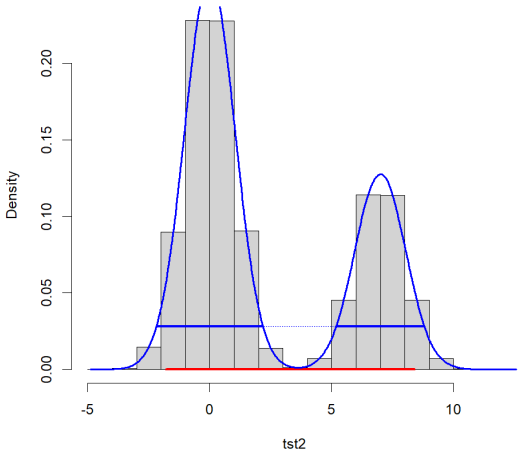
- there exists an infinite number of $(1-\alpha)$ credible regions. In single parameter case, we seek for
  - the shortest interval,
  - an equal tailed interval, i.e., $[l, u]$ such that $\Pi(\theta \leq u|x) = \Pi(\theta \geq l|x) = \alpha/2$,
  - an Highest Posterior Density (HPD) interval.

Applied to our previous example

```
# Summarize the posterior in a table:
A    <- y+a
B    <- n-y+b
Mean <- A/(A+B)
Var  <- A*B/((A+B)*(A+B)*(A+B+1))
SD   <- sqrt(Var)
alpha=0.05
Q05  <- qbeta(alpha/2,A,B)
Q95  <- qbeta(1-alpha/2,A,B)
P0.1  <- pbeta(0.1,A,B, lower.tail = TRUE)

output <- cbind(Mean,SD,Q05,Q95,P0.1)
output <- round(output,4)
output

##        Mean     SD    Q05    Q95   P0.1
## [1,] 0.2097 0.0513 0.1186 0.3184 0.0066

cat("95 % credible interval", c(qbeta(0.025, a_post, b_post), qbeta(0.975, a_post, b_post)))

## 95 % credible interval 0.1186424 0.3184212
```

Definition 1.4 (HPD region)**.** C^{π}_{α} *is an HPD region if and only if* $C^{\pi}_{\alpha} = \{\theta; \pi(\theta|x) \geq h_{\alpha}\}$*, where* h_{α} *is some value such that* $\sup_h \int_{\{\theta:\pi(\theta|x)>h\}} \pi(\theta|x)d\theta \geq 1-\alpha$*.*

Here we illustrate a very appealing feature of HPD intervals considering a bimodal distribution



'blue horizontal lines' correspond to the HPD interval; 'red horizontal line' to an equal tail credible interval.

1.2.4 Monte Carlo approximation to posterior characteristics

We are typically interested in quantities of the form

$$\mathbb{E}_{\theta|x}(g(\theta)) = \int_{\Theta} g(\theta)\pi(\theta|x)d\theta. \tag{1}$$

Since $\pi(\theta|x)$ is a density, we can use a sample $\theta_1, \ldots, \theta_M$ generated from the posterior density $\pi(\theta|x)$ and approximate (1) by an empirical average

$$\bar{g}_M := \frac{1}{M}\sum_{m=1}^{M} g(\theta_m).$$

By the SLLN, $\bar{g}_M$ converges a.s. to $\mathbb{E}_{\theta|x}(g(\theta))$.
The rate of convergence of $\bar{g}_M$ can be assess by the variance

$$\mathbb{V}ar(\bar{g}_M) = \frac{1}{M}\int_{\Theta}(g(\theta) - \bar{g}_M)^2\pi(\theta|x)d\theta$$

which is estimated from the sample $\theta_1, \ldots, \theta_M$ by

$$\hat{\mathbb{V}ar}(\bar{g}_M) = \frac{1}{M^2}\sum_{m=1}^{M}(g(\theta_m) - \bar{g}_M)^2.$$

```
M=10000
theta_post <- rbeta(n = M, a_post, b_post)
cat("mean", mean(theta_post))

## mean 0.2097799

cat("95 % credible interval", c(round(quantile(theta_post, probs = 0.025),4),
                                round(quantile(theta_post, probs = 0.975), 4)))

## 95 % credible interval 0.1169 0.3175

cat("The posterior probability that theta is smaller than 0.1 is", sum(theta_post<0.1)/M)

## The posterior probability that theta is smaller than 0.1 is 0.0073
```

1.2.5 Convergence of the posterior distribution

In a pure Bayesian approach "everything" is random, there is no "true parameter". Bayesian inference ends with exploiting the posterior. Nevertheless, we can consider an hybrid Bayesian framework under which we will study the properties of the posterior under the law P_{θ_0} from which the observations have been generated (alike in the frequentist approach). We are then interested in the **frequentist properties of the posterior** among which the consistency as defined hereafter for any $\delta > 0$:

$$\Pi(\{\theta : \|\theta - \theta_0\| > \delta\} | x_1, \dots, x_n) \to 0, \ P_{\theta_0}, \text{ as } n \to \infty.$$

Hereafter we give a more general result which basically states that Bayesian procedures (for Frequentist inference) give the same results as if we were using the asymptotic distribution of maximum likelihood estimators. Before giving this result, we give the following definition and lemma.

Definition 1.5. *Let P, Q be two probability measures with $dP = pd\mu$ and $dQ = qd\mu$. The L_1-distance between P and Q is $\|P - Q\|_1 = \int |p - q| d\mu$ with respect to some measure μ.*

Lemma 1.1. *Let P, Q be two probability measures on $\mathcal{X}$ with $\sigma-$algebra $\mathcal{A}$. Then,*

$$\|P - Q\|_1 = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Theorem 1.1 (Bernstein von-Mises)**.** *Let $\mathcal{P} = \left\{P_\theta; \theta \in \Theta \subset \mathbb{R}^d\right\}$ be a Bayesian (regular) statistical model. Let $\boldsymbol{x}$ be our sample generated from P_{θ_0}, $\theta_0 \in \Theta$. Suppose that the prior distribution on Θ satisfies*

- *Π has a density π w.r.t the Lebesgue measure on $\mathbb{R}^d$.*
- *$\pi(\theta_0) > 0$ and $\pi(.)$ is continuous at θ_0.*

Assume that the Fisher information matrix $I(\theta_0)$ is invertible at θ_0[1]. Let $\hat{\theta}^{MLE}$ be the MLE in this model. Then,

$$\left\| \Pi(\cdot|\boldsymbol{x}) - \mathcal{N}\left(\hat{\theta}^{MLE}, \frac{I(\theta_0)^{-1}}{n}\right)(\cdot) \right\|_1 \overset{P_{\theta_0}}{\to} 0, \ n \to \infty.$$

In the Bayesian setting, there is no need to estimate the Fisher information matrix which is generally in practice unknown.

Theorem 1.2. *Consider a statistical model for which the parameter space $\Theta \subset \mathbb{R}$ and that we obtain a posterior $\Pi(\cdot|x)$ from the prior Π and data $\boldsymbol{x} = (x_1, \dots, x_n)$. Let $\alpha \in (0, 1)$, z_α be the α-quantile of a standard normal. Suppose that the BvM theorem applies, then, for $l_n(x), u_n(x)$ defined as*

$$\Pi\left((-\infty, l_n)|x\right) = \alpha/2$$
$$\Pi\left((u_n, \infty)|x\right) = \alpha/2.$$

Then,

$$[l_n, u_n] = \left[\hat{\theta}^{MLE} - \frac{z_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}\left(1 + o_P(1)\right), \hat{\theta}^{MLE} + \frac{z_{1-\alpha/2}}{\sqrt{nI(\theta_0)}}\left(1 + o_P(1)\right)\right],$$

where $o_P(1)$ is an arbitrary quantity going to 0 as $n \to \infty$ under P_{θ_0}.

Lemma 1.2. *Let P, Q be two probability measures, then*

$$\|P - Q\|_1 \leq 2\sqrt{KL(P, Q)}$$

with $KL(P, Q) = \int \log(p/q) p d\mu$ the Kullback-Leibler divergence.

Lemma 1.3. *For all $\mu \in \mathbb{R}$ and $\sigma^2 > 0$,*

$$KL(\mathcal{N}(0, \sigma^2), \mathcal{N}(\mu, \sigma^2)) = \frac{\mu^2}{2\sigma^2}$$
$$KL(\mathcal{N}(0, 1), \mathcal{N}(0, \sigma^2)) = \log \sigma + \frac{1 - \sigma^2}{2\sigma^2}$$

[1]Reminder. In the univariate case, under suitable regularity conditions, the Fisher Information matrix is given by $I(\theta) = -\mathbb{E}_\theta \frac{d^2}{d\theta^2} p(x|\theta)$

Illustration Consider the normal-normal Bayesian model. The sampling family is $\{\mathcal{N}(\theta,1)\}$, the prior is $\Pi=\mathcal{N}(a,1)$ for a given $a\in\mathbb{R}$. The posterior distribution is $\mathcal{N}\left(\bar{\theta},\frac{1}{n+1}\right)$ with $\bar{\theta}=\frac{a+n\bar{X}}{n+1}$.

$$\begin{aligned}
\left\|\mathcal{N}\left(\bar{\theta},\frac{1}{n+1}\right)-\mathcal{N}\left(\bar{X},\frac{1}{n}\right)\right\|_1^2 &= \left\|\mathcal{N}\left(\bar{\theta}-\bar{X},\frac{1}{n+1}\right)-\mathcal{N}\left(0,\frac{1}{n}\right)\right\|_1^2 \\
&= \left\|\mathcal{N}\left(\sqrt{n}(\bar{\theta}-\bar{X}),\frac{n}{n+1}\right)-\mathcal{N}\left(0,1\right)\right\|_1^2 \\
&\leq 4\left\{\left\|\mathcal{N}\left(\sqrt{n}(\bar{\theta}-\bar{X}),\frac{n}{n+1}\right)-\mathcal{N}\left(0,\frac{n}{n+1}\right)\right\|_1^2+\left\|\mathcal{N}\left(0,\frac{n}{n+1}\right)-\mathcal{N}\left(0,1\right)\right\|_1^2\right\} \\
&\quad \text{using the inequality } (a+b)^2\leq 2a^2+2b^2, \\
&\leq 8\left\{\frac{1}{2}\log\left(\frac{n}{n+1}\right)+\frac{1}{2}\left[\frac{1-\frac{n}{n+1}}{\frac{n}{n+1}}\right]+\frac{n(\bar{\theta}-\bar{X})^2}{2\frac{n}{n+1}}\right\} \quad \text{(from the previous lemma)} \\
&\leq 4\log\left(1-\frac{1}{n+1}\right)+\frac{4}{n}+4(n+1)\left(\frac{a-\bar{X}}{n+1}\right)^2 \\
&\quad \text{(from the fact that } \log(1-x)\leq x \text{ for small } x) \\
&\leq -\frac{4}{n+1}+\frac{4}{n}+\frac{4}{n(n+1)}\left(\sqrt{n}(\bar{X}-a)\right)^2 \\
&= \frac{4}{n(n+1)}\left[1+\left(\sqrt{n}(\bar{X}-a)\right)^2\right]
\end{aligned}$$

Or, we remark that

$$n(\bar{X}-a)^2\leq 2n(\bar{X}-\theta_0)^2+2n(\theta_0-a)^2,$$

Since $\sqrt{n}(\bar{X}-\theta_0)\sim\mathcal{N}(0,1)$ we conclude using the continuous mapping theorem [2] that

$$\left\|\mathcal{N}\left(\bar{\theta},\frac{1}{n+1}\right)-\mathcal{N}\left(\bar{X},\frac{1}{n}\right)\right\|_1^2\leq\frac{8}{n(n+1)}\left[n(\bar{X}-\theta_0)^2\right]+\frac{4}{n(n+1)}\left[1+2n(\theta_0-a)^2\right]$$

which tends to 0 as $n\to\infty$.

1.3 Minimax estimation

In this section, we discuss how the Bayesian paradigm furnishes good frequentist estimators.

Consider a parametric statistical model $\{P_\theta,\theta\in\theta\}$. Define $\mathcal{D}$ the set of possible decisions, i.e., the set of functions from Θ to $g(\Theta)$ where $g(\cdot)$ depends on the context:

- estimation of θ, then $\mathcal{D}=\Theta$
- test $\mathcal{D}=\{0,1\}$.

We consider hereafter only the estimation context.

Definition 1.6. *A loss function is a measurable function:*

$$L:\Theta\times\mathcal{D}\to\mathbb{R}^+$$

Examples

- $L(\theta,\hat{\theta})=\left\|\theta-\hat{\theta}\right\|^2$
- $L(\theta,\hat{\theta})=\left|\theta-\hat{\theta}\right|$

Definition 1.7. *Frequentist risk:*

$$R(\theta,\hat{\theta})=\mathbb{E}_\theta L(\theta,\hat{\theta})=\int_{\mathcal{X}}L(\theta,\hat{\theta})f(x;\theta)dx$$

*this is a function of θ which does not define a **total ordering** on $\mathcal{D}$, i.e., we cannot compare all decision or estimation.*

[2] **Theorem 1.3** (Continuous mapping theorem). *Let $g:\mathbb{R}^k\to\mathbb{R}^m$ be continuous at every point of a set C such that $P(X\in C)=1$. If $X_n\xrightarrow{*}X$, then $g(X_n)\xrightarrow{*}g(X)$, where $*$ stands for d,p,as.*

Example: $(X_1, \ldots, X_n)$, $X_i|\theta \sim Be(\theta)$, $\theta \in \Theta = (0,1)$. Consider the quadratic risk, the two following estimators $\hat{\theta}_1 = \bar{X}$, $\hat{\theta}_2 = \frac{Y+\sqrt{n}/2}{n+\sqrt{n}}$, $Y = \sum_{i=1}^n X_i$.

$$R(\hat{\theta}_1, \theta) = \frac{\theta(1-\theta)}{n}$$
$$R(\hat{\theta}_2, \theta) = \frac{n}{4(n+\sqrt{n})^2}.$$

None of these estimators uniformly dominates the other. To be able to compare them, we should summarize their performance by a scalar. There are two possibilities:

- Maximum risk $\overline{R}(\hat{\theta}) = \sup_{\theta\in\Theta} R(\theta, \hat{\theta})$.
- Bayesian risk under prior π.
$$B_\pi(\hat{\theta}) = \int_\Theta R(\theta, \hat{\theta})\pi(\theta)d\theta.$$

According to the maximal risk, we prefer $\hat{\theta}_2$ to $\hat{\theta}_1$. But we see that if n is large then $\hat{\theta}_2$ is better only around $1/2$.

Definition 1.8. *Posterior Bayesian risk: Given a prior π, a loss function, the posterior Bayesian risk is given by*

$$\begin{aligned} r_\pi(\hat{\theta}|x) &= \int_\Theta L(\hat{\theta}, \theta)\pi(\theta|x)d\theta \\ &= \mathbb{E}_{\theta|x} L(\hat{\theta}, \theta). \end{aligned}$$

Theorem 1.4. *The Bayes risk satisfies*

$$B_\pi(\hat{\theta}) = \int_{\mathcal{X}} r_\pi(\hat{\theta}|x)p(x)dx.$$

Proof.

$$\begin{aligned} B_\pi(\hat{\theta}) &= \int_\Theta R(\theta, \hat{\theta})\pi(\theta)d\theta \\ &= \int_\Theta \int_{\mathcal{X}} L(\theta, \hat{\theta})f(x|\theta)dx\pi(\theta)d\theta \\ &= \int_{\mathcal{X}} \int_\Theta L(\theta, \hat{\theta})\pi(\theta|x)d\theta f(x)dx \\ &= \int_{\mathcal{X}} r_\pi(\hat{\theta}|x)p(x)dx. \end{aligned}$$

□

If $\hat{\theta}$ is the value which minimizes $r_\pi(\hat{\theta})$ then $\hat{\theta}$ is called the Bayes estimator.

Theorem 1.5. *Consider the quadratic risk, then the Bayes estimator is $\hat{\theta} = \mathbb{E}_{\theta|x}\theta$ the posterior mean. Indeed, $\|\cdot\|^2$ is convex and two times differentiable, then*

$$\frac{\partial}{\partial\hat{\theta}}\left[\int_\Theta (\hat{\theta}-\theta)^2\pi(\theta|x)d\theta\right] = 0$$

Back to our example that means that, under $Beta(\alpha, \beta)$ prior, the Bayes estimator is $\frac{y+\alpha}{n+\alpha+\beta}$.

Back to our example, we seek for an estimator which minimizes the maximal risk, i.e.,

$$\sup_{\theta\in\Theta} R(\hat{\theta}, \theta) = \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} R(\tilde{\theta}, \theta) \equiv R_n(\Theta)$$

where the infemum is taken over all estimators $\tilde{\theta}$.

We have two aims: find $\mathbb{R}_n(\Theta)$ and an estimator which attains this risk.

Sometimes it is feasible only asymptotically.

Hereafter, let us denote $a_n \sim b_n$ as $\frac{a_n}{b_n} \to 1$, $n \to \infty$

$$\sup_{\theta\in\Theta} R(\hat{\theta},\theta) \sim \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} R(\hat{\theta},\theta),\ n\to\infty.$$

Some other times,

$$\sup_{\theta\in\Theta} R(\hat{\theta},\theta) \asymp \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} R(\hat{\theta},\theta)$$

$a_n \asymp b_n$ as $\frac{a_n}{b_n}$ and $\frac{b_n}{a_n}$ are bounded. Finding such estimators can be complicated, however a Bayes estimator with a constant risk function is minimax.

Theorem 1.6. *Let $\hat{\theta}$ be an estimator. Suppose*

1. *its risk $R(\hat{\theta},\theta)$ is a constant function of θ.*
2. *it is Bayes estimator under prior π.*

Then, $\hat{\theta}$ is minimax.

Proof. By contradiction, suppose that $\hat{\theta}$ is not minimax, then there exists $\tilde{\theta}$ such that $\sup_{\theta\in\Theta} R(\theta,\tilde{\theta}) < \sup_{\theta\in\Theta} R(\theta,\hat{\theta})$, then

$$\begin{aligned} B_\pi(\tilde{\theta}) &= \int R(\theta,\tilde{\theta})\pi(\theta)d\theta \\ &\leq \sup_{\theta\in\Theta} R(\theta,\tilde{\theta}) \\ &< \sup_{\theta\in\Theta} R(\theta,\hat{\theta}) \\ &= \int R(\theta,\hat{\theta})\pi(\theta)d\theta \\ &= B_\pi(\hat{\theta}). \end{aligned}$$

This is a contradiction since $\hat{\theta}$ is the Bayes estimator for π it minimizes B_π. □

Back to our example $\hat{\theta} = \frac{\alpha+y}{n+\alpha+\beta}$, compute the risk $R(\theta,\hat{\theta})$ this is a quadratic function of θ, it is constant if and only if the coefficient of θ are 0, i.e. it is the solution of:

$$(\alpha+\beta)^2 = n\ ;\ 2\alpha(\alpha+\beta) = n.$$

That has for solution $\alpha = \beta = \sqrt{n}/2$. Hence, the minimax estimator of θ is

$$\hat{\theta}_2 = \frac{y+\sqrt{n}/2}{n+\sqrt{n}}.$$